**Motivation**

- In some scientific fields, the recent ability to collect multitudes of data easily and quickly have made computational abilities vital to researchers and practitioners. Meanwhile, fields previously thought to be niche disciplines, such as computational biology, are now “becoming an integral part of the practice of biology across all fields” [@bootcamp].

- The National Science Foundation’s *Vision and Change in Undergraduate Biology* stated, “students should be competent in communication and collaboration, as well as experience with modeling, simulation, and computational and systems-level approaches to biological discovery and analysis, and should be familiar with using large databases.”

**Computing in Statistics**

- Computing is an increasingly important and necessary aspect of a statistician’s work and needs to be incorporated more fully into statistics training.

- Duncan Lang asks us to consider if “in our courses, do students build the confidence needed to overcome computational challenges?” (Duncan Lang, 2010)

**What Do We Teach?**

- However, the 2014 ASA curriculum guide recommends that "statisticians at any level need to master an integrated combination of skills built upon statistical theory, statistical application, data management, computation, mathematics, and communication."

**Why So Rare?**

- But Temple Lang argues that it is crucial to have a basic course that addresses computational reasoning and the fundamentals of programming with data, which is very different from a typical introductory programming class (2010).

**Computational Knowledge Acquisition**

- Most researchers "learn what they know about programming and data management on their own or the information is passed down within a lab."

- This process results in substantial hidden costs:

\* "researchers spent weeks or months doing things that could be done in hours or days,

\* they do not know how trustworthy their results are, and

\* they are often unable to reproduce their own work, much less that of their colleagues" [@carpentry].

**The Lens of Distributed Cognition**

- Investigates the process by which cognitive resources are shared socially to extend an individual's ability to accomplish something that they otherwise could not achieve alone.

- Extends what is considered cognitive beyond the individual, encompassing social interactions with their peers, teachers, networks, and resources and tools in their environment.

**Beliefs of Distributed Cognition**

- Individuals coordinate different types of structure in their environment.

- Coordinating these different structures requires effort.

- Individuals off-load tasks with high cognitive effort to their environment, whenever possible.

- Dynamics of cognitive load-balancing are available through social organizations.

**Research Questions**

1. What computing skills are necessary for environmental science graduate students to successfully implement applications of statistical computing in their research?

2. How are students filling the gap between the computing skills they know and the computing skills they need to know, to perform applications or statistics in their research?

3. How can workshops help to alleviate this gap between statistical computing knowledge and expectations?

**Pilot Study**

* Describe study: five women in environmental science fields, enrolled in Stat 512 in spring of 2017.
* In our pilot study we asked environmental science graduate students to complete applications of statistical computing, in the context of ecological data.
  + These questions demonstrated the computational skills students were unfamiliar with.
  + However, we were only able to glimpse the computational skills they knew.
* Lead to RQ1: interested in what they need to know, not what they don’t know.

**Paths for Acquisition of Statistical Computing Skills**

* Themes for knowledge acquisition that developed throughout the participants’ interviews were,
  + peer support,
  + singular consultant, and
  + independent research.
* Coursework appeared within peer support and independent research.
  + However, coursework was consistently voiced to depend on peer assistance or independent research.

**Computational Support through Workshops**

- "Currently the resources for training in data-intensive research skills are both broad and scattered, complicating navigation for both novices and experts alike" (Hampton, p. 552).

- “Many online programs are user paced, so that knowledge can advance rapidly. However, these programs have drawbacks. There are initial barriers to entry, but once a basic understanding of coding has been achieved, students can more readily gain additional skills.” (Hampton, p. 555)

- In addition, students pick up bad habits, misunderstandings, and, more importantly, the wrong concepts.

- They learn just enough to get what they need done, but they do not learn the simple ways to do things nor take the time to abstract what they have learned and assimilate these generalities.

- Their initial knowledge shapes the way they think in the future and typically severely limits them, making some tasks impossible.

**Design Study**

- Broad features:

\* help to address problems that arise from practitioners' attempts to support student learning,

\* account for interventionist nature of developing instructional methods,

\* hold strong theoretical and pragmatic orientations,

\* involve iterative cycles of design and analysis,

\* aim for generalizability.

**Components of a Classroom Design Study**

- Classroom design studies consist of six stages,

\* specifying goals for students' learning,

\* documenting instructional starting points,

\* delineating an envisioned learning trajectory,

\* placing the study in a theoretical context, and

\* iterative cycles of design, data collection, and analysis.

**Design Study & Distributed Cognition**

- The tenants of distributed cognition speak to the need for an observational methodology, as no other method can illuminate "what matters" in a setting.

- As the design process creates new tools for individuals, it adds new cognitive interactions to the study.

\* This iterative process of observation to theory to design and back is an important cycle of the framework suggested for distributed cognition research.

**Learning Goals**

* It is unreasonable to expect that every researcher is an expert in their domain science, statistics, data management, processing, and visualization [@hampton, p. 547].
* Instead, researchers should \*at least\* be familiar with these concepts, to promote effective collaborations.
* Researchers should have the tools to,   
  + work with messy or organized datasets, stored in varied data formats, in a reproducible workflow,
    - familiarity with metadata generation and processing,
    - manipulate data sets into consistent formats appropriate for analysis
  + employ alternative statistical methods and simulation,
    - computational approach to statistics training
    - avoid overload of highly specialized methods, contingent on a narrow set of assumptions,
    - use scripting language for statistical analysis,
  + make use of basic software skills,
    - NOT be software developers, but be familiar with:
      * learning a computing language
      * writing clean, well-organized, read-able code
      * identify common tasks that can be separated into functions
      * discovering, assessing, and managing dependencies within software
      * learning how to find software that already provides the required functionality
      * “If you get through graduate school in a STEM field and you have had no exposure to any form of scripting or programming, I think you are handicapped. I don't care what language it's in. The ability to serialize a process and code it is really critical.” - Geoff
  + create effective data visualizations
    - create visualizations early in the data-exploration process
    - use as tools to understand relationships and/or trends that inform analysis methods
    - create visualizations that maintain a close connection with original data
    - interactive visualizations (plot.ly and Shiny) and mapping (leaflet)
* Of course, not all institutions have the ﬂexibility to develop a statistics course on computing with data, but it is crucial to have a basic course that addresses computational reasoning and the fundamentals of programming with data, which is very different from a typical introductory programming class.

**Knowledge Starting Points**

- In reasoning through computational problems, students misconceptions of foundational concepts became unveiled.

- Example: Many misunderstandings revolved around how `R` interprets logicals (TRUE/FALSE).

\* "Using the subset command where you call the data fame and you have your \[\], and just do -c(species = c("Bull", "WCT"))."

\* "I think you can remove observations by putting the observation number, but I don't know exactly how."

\* "I didn't know that was what was going on [subset using logicals]. I don't know if I was ever taught that."

**Expectations Starting Points**

* Although the environmental science community has come to an agreement on a "common set of best-of-class software tools" (Hampton, 2015), the environmental science community has not agreed upon a common set of computational skills researchers should have.
* Faculty from Environmental Science fields were interviewed regarding the computational expectations they hold for graduates of their Master's and Doctoral programs.
* Computational expectations varied between fields of research, however, many faculty emphasized:
  + writing functions,
  + using conditional statements,
  + looping & vectorization,
  + database storage, and
  + data manipulation.

**Theoretical Framework**

* It is commonly accepted that statistics should be taught in context, so why teach statistical computing in the context of solving scientiﬁc problems through data analysis?  
  + Programming is technical and often frustrating, but when embedded in exploring data, drawing plots, looking for anomalies, making conjectures, and looking for supporting evidence, the students learn the computational aspects as part of an interesting, challenging, exciting, conﬁdence-building process.
  + For most students, it is a big leap from practicing basic programming skills to embracing problem-solving methodologies and general computing principles. Students, ideally, gain this experience as they behave like scientists/statisticians who work with data, that is, when they compute with data in the context of solving a scientiﬁc problem.
  + With this approach, students are exposed to a much more subjective, creative activity than typically encountered in traditional computing and statistics methodology classes, and as a result gain a deeper, richer appreciation for the practice of statistics.
  + Statistical computing topics that are integrated with data and context add a new pedagogical dimension to the entire statistics curriculum and can expose students to methodology that they would not typically encounter.
* Students gain hands on experience with statistical concepts ﬂowing from contextual problem solving with data, and they make their own discoveries by posing and answering questions rather than solely ﬁtting models or using “this week’s lecture’s methodology” as a computing exercise.

**Computational Skills Data Collection**

* A cohort of first year environmental science graduate students were recruited in Spring 2018 to follow longitudinally thorough their program of study.
* One-on-one interviews investigating the computational skills they have used for their research and where they acquired those skills.
* Collect and analyze the code they generated for their research practices.
* What computational skills are necessary in a variety of fields?
* Where are these computational skills acquired?
* How are workshops impacting the development of computational abilities?
* Observations of common environmental science courses, which teach computing, and their associated course materials (syllabi, labs, online resources) will be analyzed.

**Workshop Data Collection**

* Workshop participants have and will be surveyed for,
  + demographic information (to separate contingent aspects of each workshop from necessary aspects), and
  + what they expect and want to learn (to confirm skills necessary for data-intensive environmental science research).
* Additionally, participants will be asked to
  + provide responses to "hands-on" applications of computational tasks (instructional techniques),
  + provide console output from workshop (identify areas of misunderstanding), and
  + evaluate workshop effectiveness (instructional techniques).
* Diagnostic assessments not only identify instructional starting points, but also help to track development of the participating students’ reasoning during the study.

**Analytical Framework (Longitudinal)**

* Interviews with longitudinal participants will be transcribed, and descriptive coding will be implemented to describe
  + the statistical computing skills participants employed in their research, and
  + how they acquired their knowledge of these concepts.
* Course materials will be analyzed to isolate,
  + computing concepts taught, and
  + instructional context used.
* Course observations will be transcribed to analyze
  + classroom discourse of computational concepts, and
  + pedagogical methods of instruction.

**Analytical Framework (Workshops)**

* Workshop materials will be kept from first implementation, in Fall 2018, to identify
  + the process of students’ learning in the workshop sessions, and
  + the aspects of the workshop learning environment that evolved over the duration of the study.
* Following each workshop session, debriefing meetings with the researchers and assistants will allow for the team to "share and debate their interpretations of workshop events."
* After each semester's workshops, the research team will meet to "outline a revised learning trajectory for the study, which takes into account the revisions made thus far" [@pcobb].

**Expected Results**

* The materials developed through these workshops will be publicly available to students, researchers, and faculty across Montana State University.
* Create new tools to enhance students' learning of core computational skills, necessary for environmental science research.
* Low barriers to entry, so that those seeking training (students and faculty) can begin their journey with a basic understanding of statistical computing.
* Inform environmental science researchers the key computational skills that are necessary for graduate students to successfully implement statistical analyses for their data.

**Benefits to Faculty**

* The pace of technological development can demand that workshops thrive outside of university curricula.
  + Capable of adapting materials rapidly, following computational and disciplinary changes.
* Can help to ease transition to integrating computing into the environmental science and Statistics curricula. Allow for pooling of resources, so that the materials needed to teach the topics are available in formats that can be quickly adapted and customized for different situations
* Understanding the importance of core skills for data-intensive environmental science research will help "facilitate the integration of training into the university" [@hampton, p. 555].